PCIe® 5.0 Protocol Update

Joe Cowan
PWG Member
Senior Systems Architect
Hewlett Packard Enterprise

Disclaimer

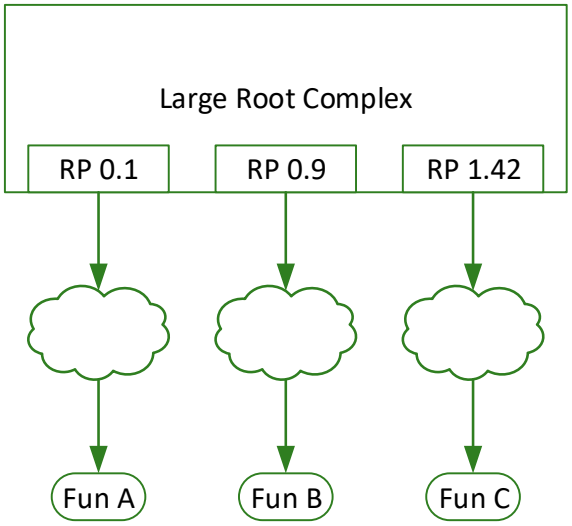
The information in this presentation refers to specifications still in the development process. This presentation reflects the current thinking of various PCI-SIG® workgroups, but all material is subject to change before the specifications are released.

Agenda

- **Completed ECNs Against the 3.1/3.1a Base Specs**
  - Hierarchy ID Message
  - Expansion ROM Validation
  - Native PCIe Enclosure Management (NPEM)
  - Link Activation
- **Selected ECRs under development**
  - Address Translation Relaxed Ordering
  - RCEC Bus Number Association
  - High-Power Slots
  - Async Hot-Plug Updates
- **Major Protocol Spec Changes for PCIe® 4.0**
- **Planned Protocol Changes for PCIe 5.0**
- **Power Subteam Introduction**

Hierarchy ID Message ECN

Hierarchy ID Problem



RP 0.x is one Hierarchy
RP 1.y is another Hierarchy

- **Bus enumeration assigns Routing IDs (RIDs) to Functions**
  - RID is unique only within a Hierarchy
- **Functions don't know which Hierarchy they are part of**
  - RID + Hierarchy ID + System ID is globally unique
  - Functions in the same Hierarchy can communicate over PCIe
- **Interesting Topologies:**
  - Large Root Complexes
  - Clustered Systems
  - Fault Tolerant Systems

Hierarchy ID Solution

- **New Type 1 Vendor Defined Message (VDM)**
  - Optional, Broadcast
  - Type 1 VDMs must be ignored by existing receivers that don't support them
- **Send from Downstream Port (ideally Root Port)**
  - Software says when to send HID Message and what HID Message contains
  - Hierarchy ID is the "Segment Group Number" defined in the Firmware Spec
  - System GUID supports multiple schemes (IEEE UID-64, Vendor Serial #, …)
- **Received by Upstream Ports**
  - Reported in capability for software/driver/firmware
- **Supports Root Complex Integrated Endpoints (RCiEP)**
  - No message on the wire – software writes RCiEP Hierarchy ID Capability
- **No operational effects**
  - Provides information to hardware / software / …
  - Does not otherwise affect PCIe operation
- **Comparison to Device Serial Number (DSN):**
  - DSN provides unique number, but usually needs a ROM
  - DSN doesn't provide location information (where am I attached)

Expansion ROM Validation ECN

Expansion ROM Validation

- **Implementation-specific methods support expansion ROM validation. This ECR does not affect those.**
- **This ECN defines a standardized mechanism to report validation results.**
  - The report is advisory
  - Does not affect access to the ROM
- **ECN defines a new 3-bit field in the Expansion ROM BAR to report status**
  - 8 encodings of Pass / Fail / InProgress / NotSupported status
- **Software is then able to take appropriate action**
  - Permits the ROM contents to be used or not
  - Contains errors/adapter

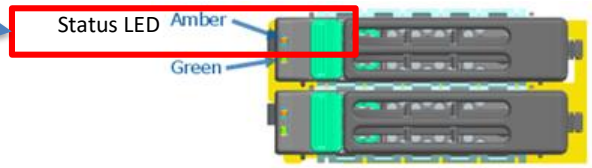
Native PCIe Enclosure Mgmt (NPEM) ECN

History and Motivation for NPEM

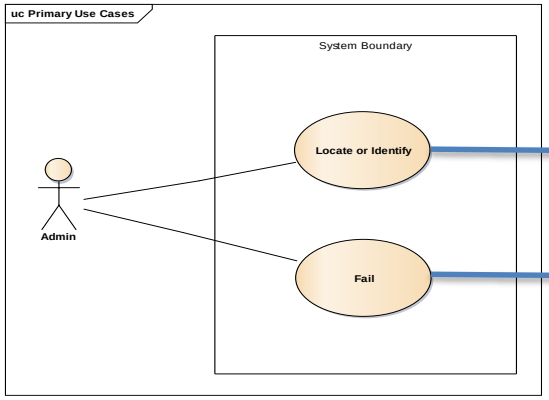
- **Disk arrays (e.g., RAID) require a visual indication of drive status (failed, critical, rebuilding, locate, etc.)**
  - Without such a visual indication, for instance, an admin can erroneously remove a good disk from the critical (degraded) RAID5 array in place of a failed disk that could lead to data loss.
- **Established enclosure/LED models not architecturally supported in PCIe**
  - SFF-8489 standard defines blink patterns for storage in accordance with the International Blinking Pattern Interpretation (IBPI).
  - SAS/SATA ecosystem defines a simple SGPIO interface for simple enclosure management (e.g., LED control)
  - A simple SGPIO unit is typically integrated to a central SAS/SATA controller/HBA.
- **Enclosure/LED function is under the purview of PCIe**
  - Unlike SAS/SATA, In NVMe, the controller/HBA is part of each drive thus the notion of a separate central controller is eliminated.
  - In typical white box server implementations for NVMe, enclosure function is inside a root port or a downstream switch port to which NVMe drive is connected. This takes the enclosure function outside of the NVMe subsystem and brings it under the purview of PCIe.
- **NPEM proposal submitted by the NVMe-MI workgroup**

NPEM defines mechanisms for storage enclosure management for NVMe SSDs, consistent with established capabilities in the storage ecosystem

NPEM (Storage LED) Use Cases



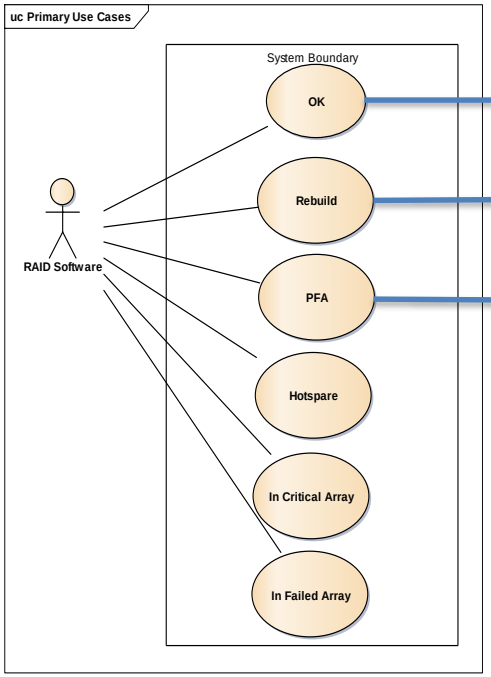
Admin initiates GUI action



Status LED blinks at 4Hz

Status LED is solid ON

RAID Software initiates action

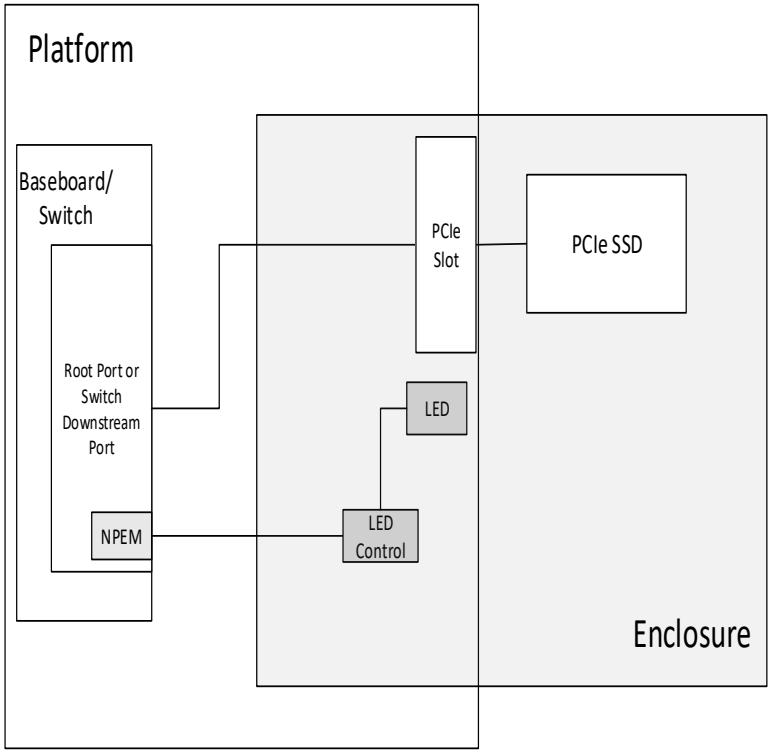


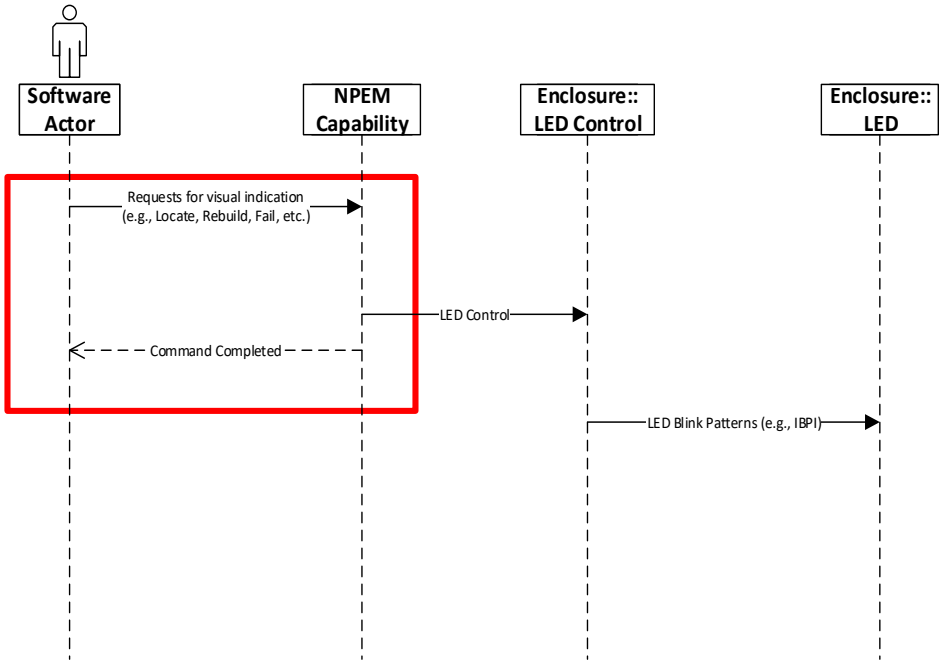
Status LED OFF

Status LED blinks at 1Hz

Status LED blinks twice at 4Hz, Pauses for 0.5s, and repeats

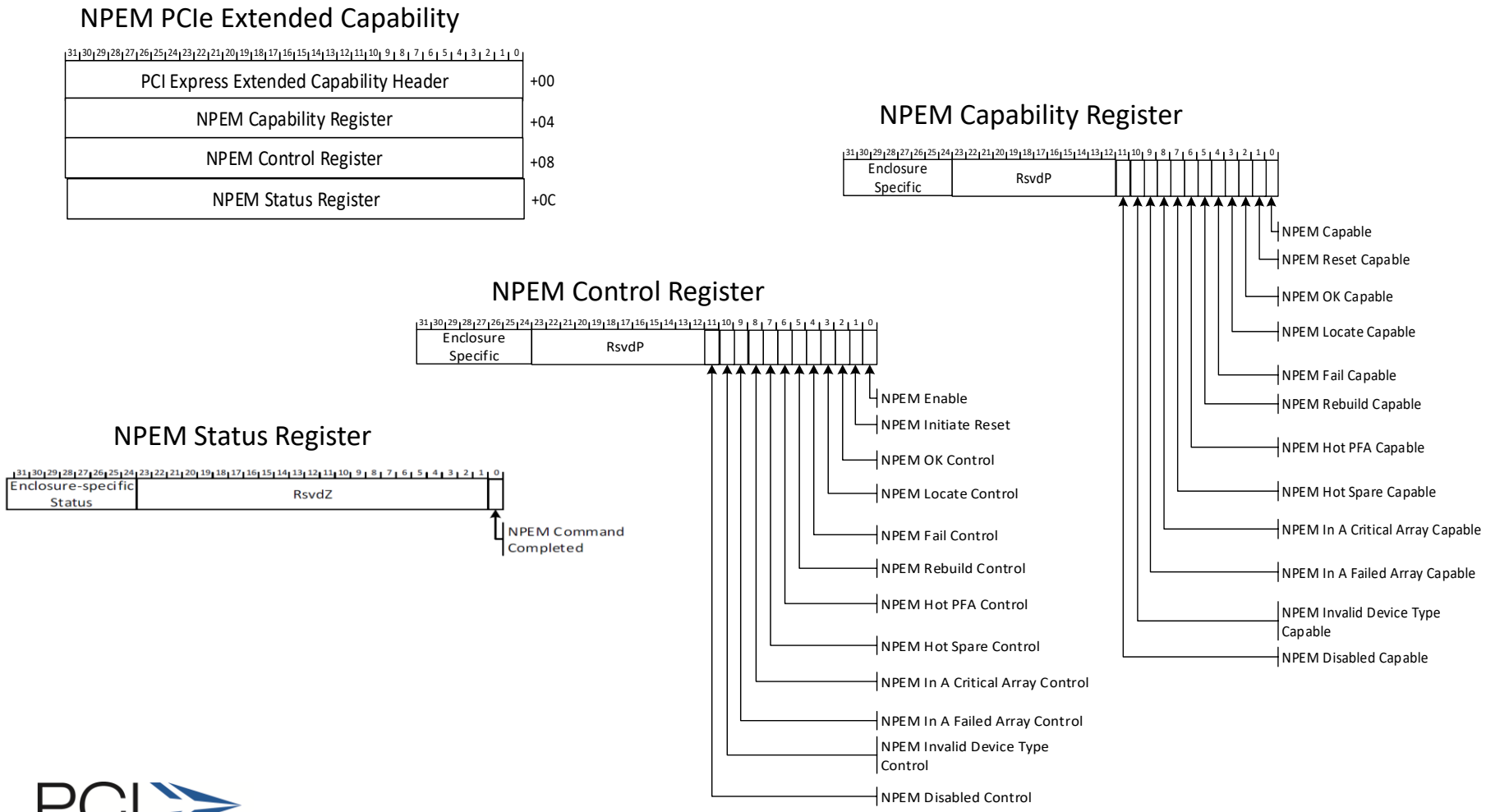
NPEM System View





NPEM provides mechanisms for enclosure management. This mechanism is designed to provide management for enclosures containing PCIe SSDs that is consistent with the established capabilities in the storage ecosystem.

NPEM PCIe Extended Capability



Link Activation ECN

Link Activation

- **Provides a mechanism for software to direct a Link out of L1.1/L1.2 before issuing PIO Request(s)**
- **Allows software to keep devices out of low-power state during latency-sensitive flows**
- **Example Use: avoid the otherwise architecturally-mandated CPU stall for a D3hot to D0 transition**

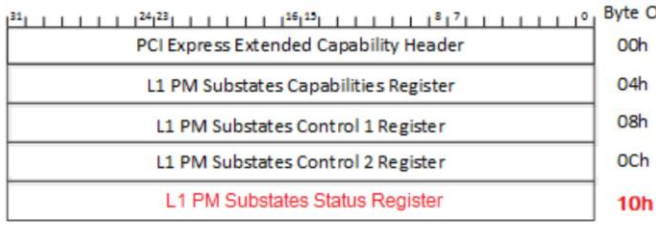


Figure 7-102: L1 PM Substates Capability

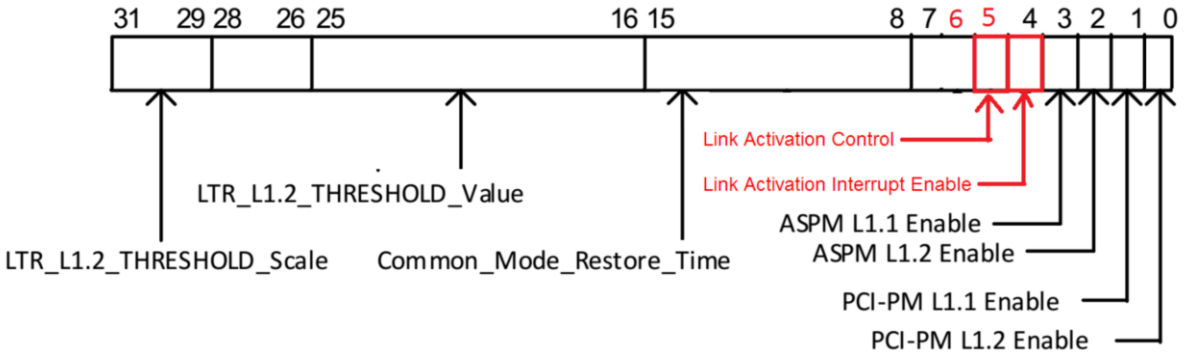


Figure 7-104: L1 PM Substates Control 1 Register

Address Translation Relaxed Ordering ECR

Address Translation Relaxed Ordering

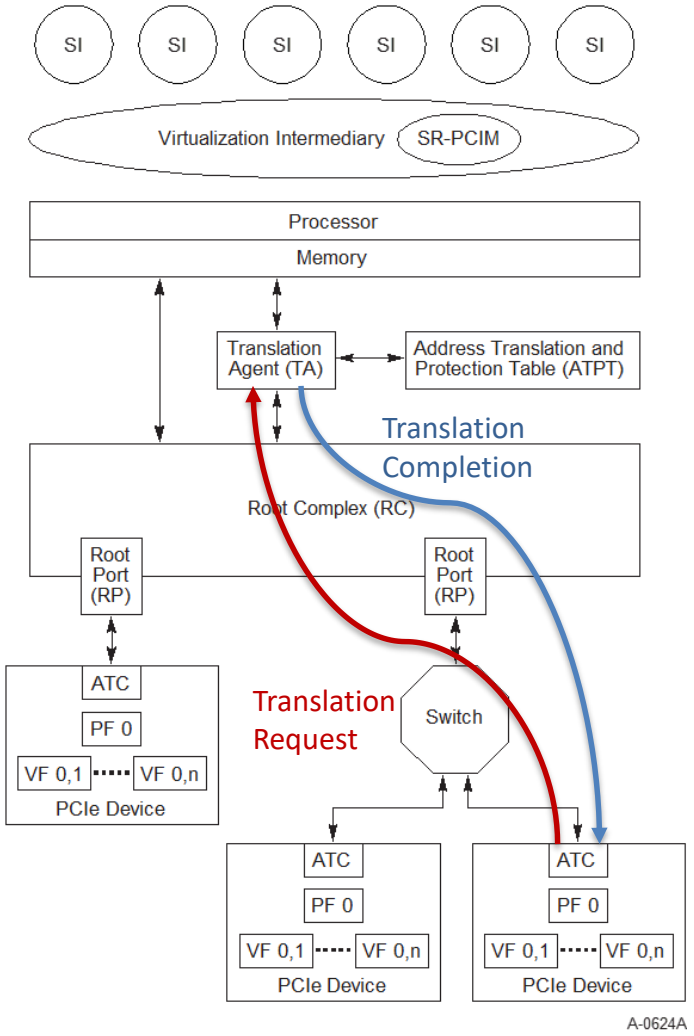
Background:

- Translation Requests and Completions are independently ordered; only needed to fill the Address Translation Cache.
  - A NP Translation Request is specified as unordered with any other request type.
  - ***However, a Translation Completion unnecessarily remains strongly ordered.***
- Normal Translated and Untranslated Requests and Completions follow normal ordering rules.
- Translation Requests & Completions do not need to be ordered with Invalidation Requests & Invalidation Completions since the ATC tracks outstanding Translation Requests.

Ordering Requirement Changes:

- Translation Requests are allowed to assert the RO bit
- TA is allowed to copy the RO bit to the Translation Completion.
- For backward compatibility, ATC must not flag an error if RO bit in the Translation Completion is 1b when it expected 0b.

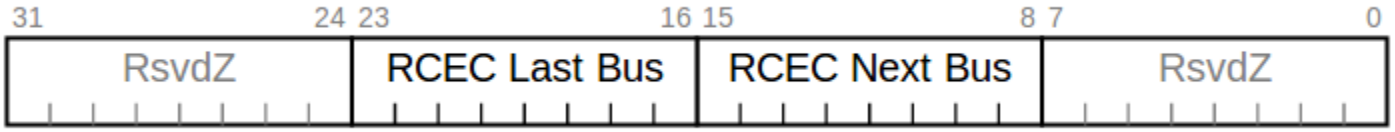
These changes allow Translation Completions to be unordered with normal traffic.



RCEC Bus Number Association ECR

RCEC Bus Number Association

- **Allows Root Complex Event Collector (RCEC) to be associated with RCiEPs on a range of Bus Numbers**
- **Simplifies support for Root Complexes with large numbers of RCiEPs**
- **Similar to Primary/Secondary/Subordinate Bus Number mechanism**
- **Adds new DWORD to RCEC Endpoint Association Capability**



High-Power Slots ECR

High-Power Slots

- **Support options up to 600 Watt**
  - Extend existing pattern to add 325, 350, … 600 encodings
  - Both Set_Slot_Power_Limit message & Power Budgeting
  - Enabling mechanism, FF specs still define supported configurations
- **Support more than 375 mA @ 3.3 V of Aux Power**
  - Hardware version of recent Firmware ECN:
    - _DSM additions for Runtime Device Power Management
  - Supports non-slot use cases (where Firmware can't help)
- **Misc cleanup**
  - Terminology, relationship between Power Budgeting and Power Mgmt capabilities…

Async Hot-Plug Updates ECR

Async Hot-Plug Updates: OS-visible HW Changes

- **Separation of inband & out-of-band Presence Detect control mechanisms**
  - Handles issues caused by IB & OOB PD always being OR'ed
- **Selection between HPS & DPC for async hot-plug**
  - HPS: long existing Hot-Plug Surprise mechanism
  - DPC: newer Downstream Port Containment mechanism
  - Enables same Downstream Port to support both HPS & DPC; SW/SFW selects which to use
- **Improved handling if add-in card is removed while Link is in low-power states like L1 or its substates**
  - Now being investigated by the PHY Logical Subteam
  - Resolution is now planned for a subsequent ECR

Async Hot-Plug Updates: Async hot-plug reference model

- **The async hot-plug reference model covers 3 areas:**
  - Async hot-plug initial configuration
  - Async removal config & interrupt handling
  - Async hot-add config & interrupt handling
- **Covers basic steps for using either HPS or DPC**
  - The HPS mech is supported but deprecated
  - New Independent OOB PD Supported feature is used
- **Covered cases:**
  - DSP supports DLL Link Active
  - Slot supports HPS, DPC, or both
  - Slots either with or without OOB PD
  - OS supports HPS and DPC; uses DPC if available
- **Reference model documented in new ImpNotes**

Async Hot-Plug Updates: Hot-Plug Intermediary (HPI) Capability

- **New functionality for Firmware-First model**
- **Block rogue Config Reads before device becomes ready following reset**
  - Avoids undefined HW behavior without requiring OS changes
- **SFW intermediary handling of PD, DLL Link Active, Device Readiness Status (DRS)**
  - Enables SFW to work around key problems without requiring OS changes
- **Hiding & configuration of newly added add-in card by System Firmware (SFW), transparent to the OS**
  - Enables SFW to handle config for security & encryption
  - Enables SFW to allow certain AICs more time to become ready

Major Protocol Spec Changes For PCIe 4.0

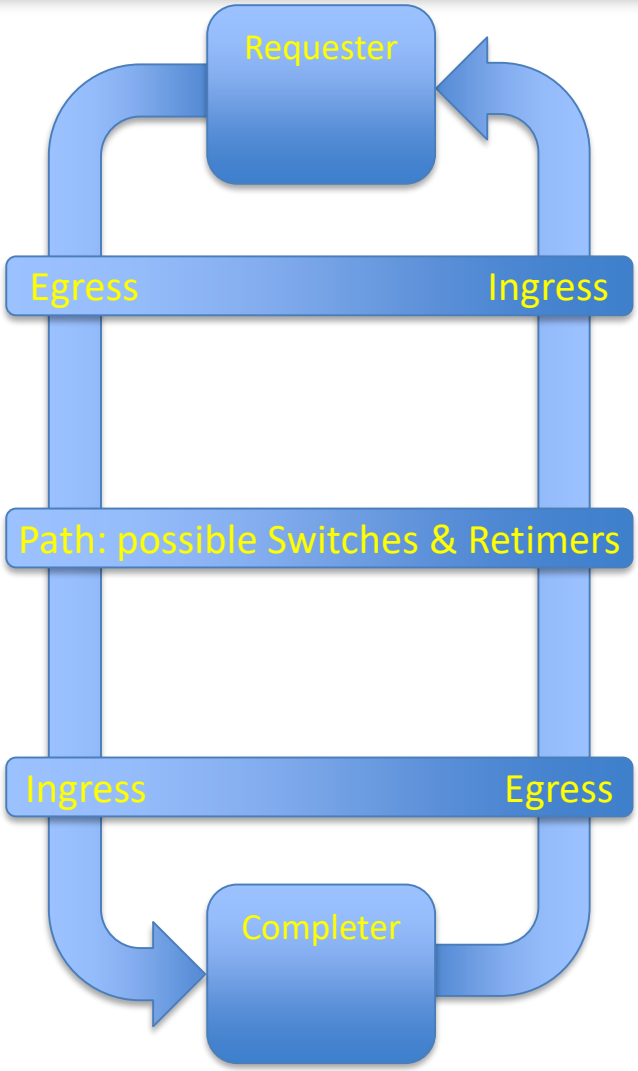
- 10-Bit Tags
- Scaled Flow Control
- Simplified Protocol Timers
- Other PCI Spec Integration

4.0 Changes: 10-Bit Tags Problem Setup

- **Max of 256 outstanding Non-Posted Requests due to 8-bit Tag field**
- **Some workloads are demanding higher numbers**
  - One important class is GPUs or GP-GPU accelerators
  - High bandwidth with relatively small transaction sizes; e.g., 64 bytes
- **Basic formula: BW = S * N / RTT with non-saturated Links, where:**
  - BW = payload bandwidth
  - S = transaction payload size
  - N = number of outstanding Non-Posted Requests (NPRs)
  - RTT = transaction round-trip time
- **Accommodating larger read access latency is highly desirable:**
  - Switch and/or Retimer topologies
  - Longer latencies due to heavy system loading
  - Higher host memory latencies with larger systems

4.0 Changes: 10-Bit Tags Increase Outstanding NPRs to 768

Outstanding 64B reads required to cover RTT latency						
	250 ns	500 ns	750 ns	1000 ns	1500 ns	2000 ns
2.5 GT/s	11	22	33	44	66	87
5.0 GT/s	22	44	66	87	131	174
8.0 GT/s	44	87	131	174	261	348
16.0 GT/s	87	174	261	348	522	696
32.0 GT/s	174	348	522	696	1044	1391



Benefit for example case with 64-byte reads

- 8-bit Tags (256 NPRs) cover light green areas
- 10-bit Tags (768 NPRs) add dark green areas
  - 256 of the 1024 values reserved for error detection

4.0 Changes: Scaled Flow Control Problem Statement

- **16GT/s bandwidth is higher**
- **16GT/s latency is mostly unchanged**
- **Existing platforms are running into limits**
  - Up to 127 Outstanding Header Credits
  - Up to 2047 Outstanding Data Credits
  - 8GT/s x16 can't deliver full Link bandwidth in some situations

- **Flow Control is independent of Link Speed**
  - Link can train to speed X and then switch to speed Y without renegotiating flow control
  - Thus, must not tie max outstanding credits to current Link Speed

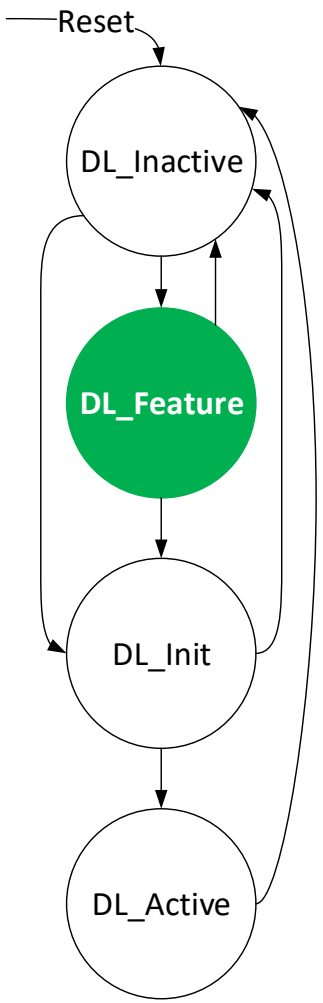
4.0 Changes: Scaled Flow Control Widen Internal Fields, Send Upper Bits

	Support	Encoding Factor	HdrFC	DLLP HdrFC:	DataFC	DLLP DataFC:
0	No	1	8 bits	HdrFC[7:0]	12 bits	DataFC[11:0]
1	Yes	1	8 bits	HdrFC[7:0]	12 bits	DataFC[11:0]
2	Yes	4	10 bits	HdrFC[9:2]	14 bits	DataFC[13:2]
3	Yes	16	12 bits	HdrFC[11:4]	16 bits	DataFC[15:4]

- **Credits don't change**
  - 1 Header Credit remains 1 TLP Header
  - 1 Data Credit remains 16 bytes
- **Support is per Link**
  - Negotiated using new Data Link Feature DLLP
- **Encoding Factor varies by credit pool**
  - {Posted, Non-Posted, Completion} × {Header, Data}
  - Receiver selects; Transmitter uses what Receiver specifies

4.0 Changes: Scaled Flow Control DL_Feature State

- **Indicates support for Scaled Flow Control**
- **Send Data Link Feature DLLP (DLF)**
  - Feature Ack bit – for handshake, initially 0b
  - Feature bits
    - Bit 0 – Scaled Flow Control Supported
    - Others reserved for future features
- **If Receive DLF**
  - Record feature bits in new Capability structure
  - Set Feature Ack in transmitted DLF
- **Exit to DL_Init if:**
  - Receive DLF with Feature Ack set
  - Receive InitFC1
- **Scaled Flow Control enabled only if both Ports support it**



4.0 Changes: Simplified Protocol Timers

- **Problem: various protocol timers had separate tables for each Link speed; i.e., one table each for 2.5GT/s, 5GT/s, & 8GT/s**
  - UpdateFC Transmission Latency Guidelines
  - REPLAY_TIMER Limits
  - Ack Transmission Latency Limit and AckFactor
  - Didn't want to add yet another set of tables for 16GT/s
- **Explored various options to permit new implementations to simplify their timer logic, while still backwards compatible**
  - UpdateFC Transmission Latency Guidelines: values for 16GT/s are identical to those for 8GT/s
  - REPLAY_TIMER Limits: developed simplified REPLAY_TIMER Limits, with % tolerances to max numbers of Symbol Times, and dramatically simplified formulas
  - Ack Transmission Latency Limit and AckFactor: simplified tables, and values for 16GT/s are identical to those for 8GT/s

4.0 Changes: Integrating Other PCI Specifications

- **Problems**
  - The PCIe Base spec was written assuming that readers have significant knowledge of PCI Local Bus specs, which is becoming less & less true
  - Some key PCIe functionality is specified only in earlier PCI Local Bus specs, which use different terminology & documentation conventions
  - A number of PCI specs are no longer being maintained
- **Solution: integrate key PCI specs into the PCIe 4.0 Base spec**
- **Specs that were integrated:**
  - *PCI™ Local Bus Specification, Revision 3.0*
  - *PCI Bus Power Management Interface Specification, Revision 1.2*
  - *Address Translation Services, Revision 1.1*
  - *Single Root I/O Virtualization and Sharing Specification Revision 1.1*

Planned Protocol Spec Changes For PCIe 5.0

- New Base Spec Development Flow
- Bypassing Equalization for Selected Speeds
- Negotiation of Alternate Link Protocols

5.0 Changes: New Base Spec Development Flow

- **Problems**
  - Word processor used for PCIe Base spec crashes frequently
  - Very difficult to have multiple editors working on the document
  - Revision control is manual with no history

- **Solution: Change to an HTML-based flow**
  - Document source converted to HTML text based format
  - Modern software development tools (Git, continuous integration, build scripts) used to manage and control spec development
    - Consistency
    - Audit trail
  - Scripts developed to produce PDFs from that source

5.0 Changes: New Protocol Features

- **Bypassing equalization for selected speeds**
  - Optionally skip EQ at 8 or 16GT/s when training to 32GT/s
    - Link not expected to operate at rates where EQ was skipped
  - Optionally not performing EQ at all if all components support it
- **Negotiation of alternate link protocols**
  - Alternate protocol: non-PCIe protocol using the PCIe PHY layer
  - Only supported with PCIe PHY with 128b/130b encoding
  - Ordered Set blocks are used as-is, including SKP Ordered Sets
  - Data block contents are determined by the alternate protocol
  - May run PCIe protocol in addition to one or more alternate protocols

Power Subteam Introduction

Power Subteam Charter

- **General responsibilities**
  - Collect various power issues in a common forum
  - Revisit legacy power material for consistency and completeness
  - Support forthcoming legal requirements (e.g., California power regulations mandating Energy-Star-like features in Desktops)
- **Relationship of this subteam with other groups**
  - A PWG subteam, not an independent WG; similar to PHY Logical
  - “Enabler” for Form Factor (FF) WGs: developing underlying mechanisms to support what they need
  - Forum to address consistency and shared nomenclature issues between FF specs, between Base and FF, etc.
  - Independent ECNs / specs not tied to Base Spec schedule

Power Subteam Efforts

- **“Big power” issues**
  - Single-slot issues like those driving the High-Power Slots ECR
  - Multi-slot issues due to scaling (e.g., 1000s of NVMe SSDs)
- **“Tiny power” issues (e.g., mobile use cases)**
- **Device-specific issues (e.g., imminent power loss)**
- **Complex issues like “white box” power discovery and negotiation**
- **Careful wording for nebulous legacy text**
- **PCI - PCIe power management schemes**
- **Errata**

Power Subteam Call to Action

- **Power affects all of us**
- **Power will never get cheaper**
- **We think with a system perspective**
- **Bring us your problems**
  - Contact us at power_subteam@pcisig.com

Thank you for attending the PCI-SIG Developers Conference 2018.

For more information, please go to www.pcisig.com

Don't forget to submit your feedback via the mobile app!

Download the **Crowd Compass** app and then search for **PCI-SIG Developers Conference** or entering the following URL into your mobile browser: https://crowd.cc/s/1rKy0
Enter event code: **DevCon2018**

Alternatively, access here: https://crowd.cc/pcisig2018

Note: Create an account within the app so Admin knows who to contact if selected as the prize winner.

Each session feedback is provided is equivalent to 1 raffle entry (up to 11 sessions).
General survey feedback = 1 raffle entry.